Final Project- Appl. Maths conc. For Deep Learning

**Deep Learning for Toxic Comment Detection**

By- Pranali Karande(101471932) Isha Jayswal(101510506)

**Data Preprocessing:**A screenshot of a computer

Description automatically generated

**Text Vectorization:**

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**Model Training:**

**A screenshot of a computer program

Description automatically generated**

A screenshot of a computer

Description automatically generated

**Making Predictions:**



A screenshot of a computer

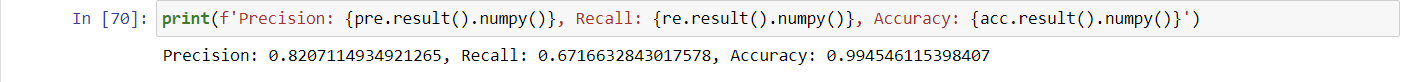
Description automatically generated

**Model Evaluation:**

A screenshot of a computer

Description automatically generated

**Training Set Results:**



**Testing the Model:A screenshot of a computer

Description automatically generated**

A screenshot of a computer code

Description automatically generated

**Report:**

This project endeavors to address the pervasive issue of toxic behavior in online discussions through the development of a deep learning model capable of automatically detecting and classifying toxic comments. Toxic behavior, encompassing hate speech, insults, and threats, poses significant challenges to fostering respectful and safe online environments. Leveraging natural language processing (NLP) techniques and deep learning algorithms, the project aims to mitigate these challenges by accurately identifying and categorizing toxic comments.

Beginning with data preprocessing, the project cleans and tokenizes a labeled dataset of comments to prepare it for model training. The model architecture, a Sequential neural network, incorporates an embedding layer to convert text data into numerical vectors, a bidirectional LSTM layer to capture sequential dependencies, and multiple dense layers for classification. Trained on annotated comments with toxicity labels across categories such as toxicity, severe toxicity, obscenity, threat, insult, and identity hate, the model undergoes evaluation using standard metrics like precision, recall, and accuracy to gauge its effectiveness in correctly classifying toxic comments.

Despite training for only a single epoch, the model demonstrates promising performance, achieving high precision, recall, and accuracy scores across various toxicity categories. Further potential for improvement is evident with additional training epochs. Moreover, the model's integration into a user-friendly interface using Gradio enables real-time interaction and toxicity predictions, enhancing accessibility and usability.

The deep learning model architecture comprises several layers, including an embedding layer, bidirectional LSTM layer, and multiple dense layers for classification. Trained on a dataset containing toxic comments labeled across various categories, the model exhibits a total of 64,091,686 parameters. During training, it achieves a loss of 0.0622 and a validation loss of 0.0457, indicative of its performance in minimizing discrepancies between actual and predicted values.

Evaluation of the model's predictions for toxicity categories reveals impressive precision, recall, and accuracy metrics, showcasing its ability to accurately classify toxic comments. Individual predictions offer insights into the model's confidence levels for each category, facilitating an understanding of its decision-making process.

In conclusion, the results underscore the efficacy of the deep learning model in detecting and classifying toxic comments, thereby contributing to the cultivation of healthier online communities. Further refinement and optimization of the model architecture hold promise for enhancing its performance and usability in real-world applications.